BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 2: e25415
doi: 10.3897/biss.2.25415

OPEN ACCESS

Conference Abstract

Service-based information extraction from herbarium specimens

Fabian Reimeier[‡], Dominik Röpert[‡], Anton Güntsch[‡], Agnes Kirchhoff[‡], Walter G. Berendsohn[‡]

‡ Freie Universität Berlin, Berlin, Germany

Corresponding author: Fabian Reimeier (f.reimeier@bgbm.org), Dominik Röpert (d.roepert@bgbm.org)

Received: 01 Apr 2018 | Published: 21 May 2018

Citation: Reimeier F, Röpert D, Güntsch A, Kirchhoff A, Berendsohn W (2018) Service-based information extraction from herbarium specimens. Biodiversity Information Science and Standards 2: e25415. https://doi.org/10.3897/biss.2.25415

Abstract

On herbarium sheets, data elements such as plant name, collection site, collector, barcode and accession number are found mostly on labels glued to the sheet. The data are thus visible on specimen images. With continuously improving technologies for collection mass-digitisation it has become easier and easier to produce high quality images of herbarium sheets and in the last few years herbarium collections worldwide have started to digitize specimens on an industrial scale (Tegelberg et al. 2014). To use the label data contained in these massive numbers of images, they have to be captured and databased. Currently, manual data entry prevails and forms the principal cost and time limitation in the digitization process. The StanDAP-Herb Project has developed a standard process for (semi-) automatic detection of data on herbarium sheets. This is a formal extensible workflow integrating a wide range of automated specimen image analysis services, used to replace time-consuming manual data input as far as possible. We have created web-services for OCR (Optical Character Recognition); for identifying regions of interest in specimen images and for the context-sensitive extraction of information from text recognized by OCR. We implemented the workflow as an extension of the OpenRefine platform (Verborgh and De Wilde 2013).

Keywords

herbarium sheets; Optical Character Recognition; image analysis; workflow; web service

Presenting author

Fabian Reimeier

Funding program

German Research Foundation (DFG)

Grant title

Ein prozessoptimiertes Standardverfahren zur Erschließung von digitalen Herbarbelegen (project number: BE 2283/12-1, STE 1635/1-1, US 118/1-1)

References

- Tegelberg R, Mononen T, Saarenmaa H (2014) High-performance digitization of natural history collections: Automated imaging lines for herbarium and insect specimens. Taxon 63 (6): 1307-1313. https://doi.org/10.12705/636.13
- Verborgh R, De Wilde M (2013) Using OpenRefine. Packt Publishing, Birmingham. [ISBN ISBN 9781783289080]